大数据为安全生产保驾护航

摘　要：大数据时代的到来，各类信息快速传播，利用海量数据和处理海量数据规范化为各行业带来了巨大的收益，促进了经济社会的快速发展。安全生产与经济社会发展密切相关，提升大数据技术在安全生产领域的应用能力至关重要。

关键词：大数据；安全生产；安全生产技术
中图分类号：F272　　文献标识码：A
文章编号：1671-0134（2018）02-085-03　　DOI：10.19483/j.cnki.11-4653/n.2018.02.033

文 / 张洪福

引言

所谓安全生产，是指在生产经营活动中，为了避免造成人员伤害和财产损失的事故而采取相应的事故预防和控制措施，使生产过程在符合规定的条件下进行，以保证从业人员的人身安全与健康，设备和设施免受损坏，环境免遭破坏，保证生产经营活动得以顺利进行的相关活动。最近几年，许多生产企业将大数据应用到自身的经营管理之中，重视大数据在安全生产中的应用价值。

1. 大数据对安全生产的影响

《中国安全生产报》2001 年 10 月 11 日创刊，是国内安全生产领域唯一综合性报纸，是传递党中央、国务院、国家安全生产监督管理局、各行业主管部门、各地方政府对安全生产工作各个阶段工作部署的重要媒介；是安全生产专业信息咨询和交流的权威平台和安全生产理论探寻、安全文化建设的主阵地；是各级安监干部工作的良师益友。作为安全生产领域权威主流媒体有着深远的影响力，能够汇聚行业内的各种数据资源，数据资源包括：各地记者站稿件、民众投稿、专家约稿、企业安全生产数据、政府安全监管数据、调查报告、安全生产相关法律知识、安全生产管理知识、安全生产技术等。作为大数据而言，除了内部数据积累，还应充分利用互联网数据，结合大数据手段对安全生产领域信息快速抓取和分析。完善生产中的数据与资料，从大数据中不断探索其中规律。

同时，2015 年 4 月 2 日，国务院办公厅印发《国务院办公厅关于加强安全生产监管执法的通知》，通知指出，要大力提升安全生产“大数据”利用能力，加强安全生产周期性、关联性等特征分析，做到检索查询即时便捷、归纳分析系统科学，实现来源可查、去向可追、责任可究、规律可循。中国安全生产报社发挥自身优势，利用大数据技术开展安全生产工作，应用价值在多个方面都能够有所体现。首先是对安全生产领域监察的敏感性强，分析基础数据可知哪些安全生产行业或某个安全生产行业哪个环节易发生安全问题。其次是有利于安全生产领域相关政策制定。中国安全生产报社拥有大量的数据支撑，对基础数据内容进行分析，便于对多因素影响下事态的发展以及在趋势方式下制定最适宜的安全举措。最后是有利于整个安全生产领域的管理推进和实施。中国安全生产报社经过有效处理海量的基础性数据，对如何安全管理已有系统性的研究。

2. 基于大数据助力安全生产

2.1 大数据积累：准确、全面地收集数据是大数据的基础

首先要充分利用已有数据。包括：各地记者站稿件、民众投稿、专家约稿、企业安全生产数据、政府安全监管数据、调查报告、安全生产相关法律知识、安全生产管理知识、安全生产技术等。

其次是充分利用互联网数据。随着网络应用技术的发展，网络信息呈现出一定的“异构”特点。网络信息仍以 HTTP 为网络传输协议，以 HTML 为展示格式，但随着互联网社区化的发展和 Web 2.0 的崛起，网页所蕴含的内容发生了深刻的变化。原来以网站 / 网页内容为主导的互联网，逐渐演变为网站、论坛（社区）、博客、微博等信息共存的局面。微信、论坛、博客、微博上蕴含的大量信息已经成为互联网上重要的信息组成部分。网民们可以在这些自媒体平台随时随地发表他们所见所闻的安全生产事件或对某个安全生产事件的态度看法等。这些自媒体平台互动性强，信息传播快，俨然成为一个舆论放大器。而且对安全生产领域来说，论坛、微博、微信上的信息比普通网站上的信息具有更重要的使用价值。安全生产事故，如燃气爆炸、坍塌事故、火灾、沉船、重大车祸等信息，都是通过论坛、微博、微信等渠道第一时间传播的。另外，一些安全生产隐患，如煤气泄漏、安全漏洞、火灾隐患等，网民可以通过互动的形式告知安监总局、安全生产报社等单位，在事故发生之前及时处理，减少人民生命财产损失，具有重大意义。

安全生产大数据的要求是对互联网上的有效信息进行采集和利用，但目前的数据采集技术主要是面向网站和网页的收集和采集，不能有效解决论坛（社区）、博客、

微博、微博的采集和更新问题。对于安全生产大数据来说，最终建设的应该是全面的信息收集机制，有效信息遍布于论坛、博客、微博、微信等载体上。针对安全生产行业特点和业务领域，选择神华集团、中石油、中石化等同类企业或同行企业的安全生产事件进行素材的收集（如央国企新闻发言稿等），历年全国各地发生的安全生产事故信息等。主要包括过往案例、对外宣传稿、分析报告等，按照事故命名、发生时间、地点、程度级别、事故类型、伤亡人数、死亡人数等属性特征进行分类，并可设定相关报道的媒体范围，同时采集与事故相关的互联网信息，形成安全生产大数据的数据支撑。

据国家安全生产监督管理总局官网数据显示，2017 年 1~7 月，全国共发生各类生产安全事故 27478 起，死亡 19783 人。其中，较大事故 377 起，死亡 1442 人；重大事故 17 起，死亡 225 人，同比增加 1 起等。及时获取这些信息，有利于相关部门了解事件态势，尽早合理决策，避免不良影响扩大化。

2.2 自然语言处理：让机器更懂人类，提高关联性等特征分析

随着人工智能的大热，国内各大企业开始纷纷布局人工智能领域，并打造出各种不同的智能终端，比如人工机器人、无人驾驶汽车、智能电视、智能冰箱……这些智能终端有一个共同的特点——不但能读懂人类语言，还能与人类交流，同时，还能进一步完成人类所下达的指令。

如此神奇的技术是如何实现的呢？这要归功于人工智能领域一项核心的处理技术——NLP。NLP（Natural Language Processing），即自然语言处理，它是研究人与计算机交互的语言问题的一门学科，也是人工智能一个重要的子领域。简单来说，NLP 是让机器“理解”人们使用的自然语言结构和意思，将自然语言翻译为机器语言形式，并加工它（总结、句法分析等），再返回给用户自然语言。它涉及很多内容和技术，如文本朗读 / 语音合成、语音识别、中文自动分词、词性标注、句法分析、自然语言生成、文本分类、信息检索、信息抽取、文字校对、问答系统、机器翻译、自动摘要、文字蕴涵……

在人工智能发展之初，NLP 技术就已经显示出巨大的魅力。1949 年埃德蒙・伯克利（Edmund Berkeley）在他出版的《Giant Brains Or Machines That Think》一书中曾写道：“最近出现许多消息，谈论的主题是奇怪的巨型机器处理信息，速度极快，技能很强……这种机器与大脑相似，由硬件和线缆组成，而不是血肉和神经，机器可以处理信息，可以计算、可以得出结论，可以选择，还可以根据信息执行合理操作。总之，这台机器可以思考。”

作为人工智能核心技术之一，自然语言处理技术越发受到技术公司的青睐，在国务院印发的《新一代人工智能发展规划》中，自然语言处理技术被列为关键共性技术。

先进的技术需要与行业进行深度结合，才能实现更大的价值。自然语言处理技术可以实现对安全生产大数据的分析处理，建立符合行业特色的安全生产知识库，包括安全生产案例库、安全生产口径库、关键词库、媒体库及敏感词库、专业领域知识库等，形成知识的积累。

安全生产案例库：首先，利用采集的行业数据，经过自然语言的解析和整理，自动从大规模行业语料中挖掘专业术语和新词，快速构建行业词典，构建行业语料库。同时，通过多个行业语料库的采样和综合，构建通用语料库。语料预处理中对语料分块，并进行分词、命名实体识别，然后进行串频统计、子串归并操作，再分别通过横向对比和纵向递进的方法进行行业术语和行业短语挖掘。可实现数据内容过滤，多语种识别和自动转码、自动分词、自动分类、自动聚类、自动热点发现、相似检索、文章排重、自动摘要、重点信息抽取等功能。案例库本着科学、实用的原则，对每个安全生产事件的特征都进行了全方位的剖析，既包括该事件的发展演变过程、网上民意演变过程图表，也包括在事件过程的各个阶段中网络上各种不同观点、看法的所占比重和典型观点的摘编。可按照事故命名、发生时间、地点、程度级别、事故类型、伤亡人数、死亡人数等属性特征进行分类，并可设定相关报道的媒体范围。业务人员可通过安全生产案例库浏览、查询和下载案例报告，利用过往的应对经验，并结合当前实际情况，提高安全生产应对处置能力。安全生产案例库是长期研究、分析互联网及行业数据积累下来的宝贵资料，对安全生产的宣传、调研、理论、培训等有一定的参考和借鉴价值。

安全生产口径库：通过自然语言处理技术，可为安全生产口径库提供技术支撑，收集并分类细分历年全国发生的安全生产事故，采集相关的媒体报道，实现提取涉事人名、地名和机构名称的功能，同时可自动标识是否涉及国务院、安监总局或各地安监局，便于分析整理各级监管机构、涉事企业及其他相关部门的处理意见、回应的时间节点、回应内容、处置方法等。可以及时、全面、准确地掌握各种信息和网络动向，从浩瀚的数据宇宙中发掘事件苗头、归纳舆论观点倾向、掌握公众态度情绪，并结合历史类似事件进行趋势预测和应对建议。建立完善的地区、机构、行业、社情民意的分类体系，便于进行信息共享、分析处理、信息快速查找，逐步形成围绕安全生产的口径知识库。通过安全生产口径库的建设，利于安全生产业务人员熟悉掌握政策、口径、提升自身业务素质，也有利于加强新闻宣传工作的组织规范性和整体协作效率，降低信息搜索成本，提高信息回应的针对性、准确性、一致性和及时性。

2.3 智能语义检索：做到检索查询及时便捷

以自然语言理解技术为基础的新一代搜索引擎，被称为智能语义检索。由于它将信息检索从目前基于关键词层面提高到基于知识（或概念）层面，对知识有一定的理解与处理能力，能够实现分词技术、同义词技术、概念搜索、短语识别以及机器翻译技术等，因而这种搜索引擎具有信息服务的智能化、人性化特征。这种允许网民采用自然语言进行信息检索，将为他们提供更方便、更确切的搜索服务。

安全生产行业搜索利用智能语义检索，能够比通用搜索提供更多的行业相关查询方式。行业搜索应提供丰富的查询手段，包括自动分类、普通检索、组合检索、拼音检索、相关短语检索等。智能语义检索更加人性化，功能也更强，能够满足行业的特殊需求。在搜索应用开发过程中，逐步选择适合于行业应用的查询方式。

2.3.1 拼音检索

拼音检索的主要功能是提供全拼检索、简拼检索、同音检索等技术，帮助用户快速有效地检索自己所需要的内容。

基于串频统计和上下文的注音技术：在大量拼音语料基础上，统计汉字串和拼音串的分布规律等大量有用信息，利用基于上下文的注音算法对多音汉字进行注音，保证了注音的准确性。

同音检索技术：支持同音检索、全拼检索和简拼检索，在丰富的拼音语料库基础上，对汉字串的分布频率进行了统计，整理出高频汉字串和拼音串的对应表，在此基础上，保证用户输入的拼音串对应的一定是最可能的汉字串。

拼音输入校正技术：利用拼音词典和相关算法实现输入校正。

2.3.2 相关短语检索

相关短语检索的主要功能是，在检索过程中，根据用户输入查询，提供一组比较常用的相关查询供用户参考，向用户提供高质量的“查询建议”，方便用户使用搜索系统。例如，输入“知识”，提示“知识管理”“知识在线”“知识经济”等。

3. 实现短语检索的关键

一是相关短语匹配技术。如何定义并计算短语的相关性是个很有挑战性的问题。相关短语匹配技术采用了语义词典和短语语法结构相结合的方法，计算短语之间的相关性，取得了满意的效果。

二是相关短语词典。相关短语词典是相关短语检索的基础，来源主要有两部分：一部分是人工整理的短语相关知识；另一部分是通过数据挖掘技术，从搜索引擎查询日志中获取的相关短语。这样既保证了词典的规模，又保证了词典的质量。新华搜索前期的工作已经形成了包含数十万条词条及其相关短语的短语词典。

三是高频查询词典。主要来源是在长期积累的检索日志基础上，整理并统计用户在日常检索中经常使用的100多万个查询。如果用户输入的短语不包含在相关短语词典中，则使用相关短语匹配技术从高频查询词典中检索相似短语。

四是人工整理和数据挖掘方法结合构造相关短语词典。相关短语词典的规模达到百万级词条和它们的相关短语，其来源主要有两部分：一部分是人工整理的短语相关知识；另一部分是通过数据挖掘技术，从搜索引擎查询日志中获取的相关短语。这样既保证了词典的规模，又保证了词典的质量。

如何定义并计算短语的相关性是个很有挑战性的问题。 实验证明了采用语义词典和短语语法结构相结合的方法，计算短语之间的相关性，取得了满意的效果。

3.1 检索结果排序

行业搜索的检索结果排序方法是研究的一个重点。通用搜索引擎采用以链接分析为主要手段的排序手段，行业搜索的检索结果排序需要综合考虑网页内容的相关性（用户查询词与网页内容的相关度）、网页自身的重要性（链接分析）以及时效性。

3.2 内容相关性：向量空间模型

传统 IR 技术中判断查询条件与文档的内容相关性，最为通用的方法是采用向量空间模型（VSM）进行计算。

安全生产智能语义检索将综合运用相关性排序、网页权重、时间权重等多种排序因素，获得较优的排序结果，具体排序过程主要基于以下与相关度相关的因素进行。

3.3 内容相关度：基于传统的 IR 排序算法

比如 TF-IDF，VSM，计算查询条件与网页的内容相关度。在网页内容方面，标题中的关键词、黑体的关键词以及标题中出现的关键词、网页外部链接的锚文本等，比网页本身内容具有更高的权重。

文档权重：主要基于链接分析方法（如 PageRank）计算文档的权重。

时间权重：按照网页发布时间（如果获取不到发布时间则取收录时间）计算时间权重。

结果排序算法的主要流程是，系统依据内容相关性、文档权重、时间权重，计算获得排序结果。

以上大数据的基础、技术和应用为大数据在安全生产中的应用提供了方向。安全生产基于大数据技术可以做到安全生产检索查询即时便捷、归纳分析系统科学。

通过对安全生产行业相关数据采集、自然语言处理、检索，可以实现资源共享、内容创新、信息增值及优质服务；通过大数据技术，逐步打造面向“互联网 +”语境下的现代化信息系统，能够充分贯彻《国务院办公厅关于加强安全生产监管执法的通知》的精神。中国安全生产报社将大数据技术与安全生产业务相结合，为安全生产领域今后的进一步改革和发展打下坚实的技术基础。媒

（作者单位：《中国安全生产报》社）